Spring 1-1-2012

Individual Differences in the Generation of Language-Related ERPs

Leif Dakota Oines
University of Colorado at Boulder, leif.oines@colorado.edu

Individual Differences in the Generation of Language-Related ERPs

by

Leif Oines,

B.A., University of California at Santa Cruz, 2008

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirement for the degree of
Master of Arts
Department of Psychology and Neuroscience
2012

This thesis entitled:

Individual Differences in the Generation of Language-Related ERPs

written by Leif Oines

has been approved for the Department of Psychology and Neuroscience

Committee Members:

_____

Albert Kim (Chair)

_____

Akira Miyake

_____

Tim Curran

Date: April 23rd 2012

The final copy of this thesis has been examined by the signatories, and we
Find that both the content and the form meet acceptable presentation standards
Of scholarly work in the above mentioned discipline.

IRB protocol # 0607.15

Oines, Leif (M.A., Psychology & Neuroscience)
Individual Differences in the Generation of Language-Related ERPs
Thesis Directed by Assistant Professor Albert Kim

Abstract

Using Event Related Potentials (ERPs), we investigated how the elicitation of two important language-related ERP components, the N400 and P600, varied across a sample of University undergraduates. Using stimuli adapted from Kim and Osterhout (2005), we examined whether subjects varied in the ERPs they generated while reading sentences containing animacy violations ("The dusty tabletops were *devouring...*") as well as similar sentences containing so-called "semantic attraction" situations ("the hearty meal was *devouring...*") in which syntactic cues indicated that the subject noun was the agent of the sentence, while semantic cues indicated that the subject would be a better fit for the theme role. Replicating Kim & Osterhout (2005), we observed an N400 in the grand-average ERPs for sentences containing animacy violations, and a P600 for sentences containing semantic attraction situations. We then examined the extent to which individual subjects conformed to this grand-average pattern by developing a measure of the tendency of each subject to show either a P600 or N400 in response to each type of sentence. By correlating this measure with subjects performance on a number of behavioral tasks, we found that participants scoring higher on a task of Verbal Working Memory Updating were more likely to show a P600 in response to animacy violations, while lower scoring subjects tended to show an N400 effect. We did not find any relationship between our behavioral variables and ERPs recorded in response to semantic attraction situations. We discuss how the ability to update verbal WM may be crucial for both syntactic *and* semantic processing of a sentence, and conclude that better updaters are more capable of updating both types of information when encountering unexpected situations such as animacy violations.

CONTENTS

I. INTRODUCTION

A growing body of work that applies electrophysiological methods to the study of language processing has afforded researchers a glimpse into brain processes that unfold in the milliseconds after listeners or readers encounter a new word in the midst of an unfolding discourse or narrative. Among the many methods available to analyze electroencephalographic (EEG) data, psycholinguists have most often employed the Event-Related Potential (ERP) to analyze changes in the EEG resulting from manipulations of lingustic stimuli. ERPs are representations of EEG data that are typically obtained by averaging epochs of EEG that immediately follow the onset of similar types of stimuli. Once many individual subject ERPs are averaged together, clear "components" in the ERP can be identified that vary reliably in amplitude in response to experimental manipulations.

ERP studies of language processing are often concerned with how manipulations of semantic and structural (i.e. syntactic) aspects of linguistic stimuli differentially effect changes to a variety of ERP components. A paradigmatic example is the work of Kutas & Hillyard (1980), which identified a clear negative deflection in the ERP about 400 ms after the onset of a visual or auditory word, the so-called N400 effect, when comparing the emphasized words in sentences such as "He spread the warm toast with SOCKS" and "He spread the warm toast with BUTTER". Replicated many times since, the N400 effect has been interpreted as indexing the ease with which an individual can integrate a word's meaning into their larger representation of the discourse up until that point, and thus is often employed in research concerned with how the brain implements semantic processing.

Other work, including Osterhout and Holcomb (1995), has identified another ERP component, the so-called P600, which has been interpreted as especially sensitive to structural aspects of language processing. When the authors compared ERPs elicited by grammatically anomalous words like "hopes" in sentences such as "The elected officials HOPES to succeed" to well-formed control versions of the

same sentences, the anomalous condition generated a significant positive deflection (compared to the control condition) in the ERP about 600 ms after word onset. Since the effect appeared to be due to the failure of the verb to grammatically agree with the number of the subject noun ("officials"), many initially interpreted the P600 as selectively sensitive to non-semantic grammatical cues conveyed by a word or sentence.

The apparent selectivity of the N400 to semantic manipulations and the P600 to syntactic factors at first seems to confirm a long-held belief by some researchers that semantic and structural processing are independent, and that both types of representations are handled by very different processes. Such a view predicts, for instance, that manipulations of semantic variables should not modulate P600 effect sizes. While there have been a number of results consistent with this prediction, several recent studies have reported data that seem irreconcilable with the idea that the P600 component is not influenced by semantic variables.

One such study by Kim and Osterhout (2005), of central importance to our research here, reported a P600 effect in response to what is arguably a clear semantic anomaly. Sentences representing their three conditions are reproduced below.

(SA1) The hearty meal was devoured by the boys	*Control*
(SA2) The hearty meal was devouring by the boys	*"Semantic Attraction" Condition*
(SA3) The dusty tabletop was devouring by the boys	*No Attraction Condition*

The principle manipulation, represented by (SA2) above, involved sentences in which an inactive subject ("meal") preceded a verb ("devouring") that would be highly plausible (in semantic terms) if its inflection was compatible with a passive structure (such as in (SA1) above). By changing the inflection of the verb to be active, a *semantically* anomalous situation was created in which the

structure of the sentence dictated that subject is acting as the agent – rather than the theme – of the verb. Crucially, this condition elicited a robust P600 effect *despite the fact that the sentence contained no grammatical anomaly.* That the P600 was driven by the "semantic attraction" between the subject noun and verb was further supported by the fact that sentences in the "no attraction" condition elicited no P600, and instead evoked the N400 effect that is usually observed in response to semantic processing difficulties. Kim and Osterhout (2005) concluded that sentences in the semantic attraction condition represent contexts in which an alternative analysis of the sentence – one in which the subject noun is assigned to the theme of verb – is considered in the face of contradicting structural cues.

A guiding hypothesis for the following research is that "semantic attraction" situations represent a linguistic "tipping point" in which subjects could pursue an interpretation that is either (1) consistent with grammatical cues – but semantically anomalous – or (2) *in*consistent with grammatical cues, but semantically coherent. Presumably, following option (1) should result in a P600, since grammatical cues would then contradict the final interpretation of the sentence. Pursuing option (2), on the other hand, should result in an N400, since "meal" is not a good agent for "devour". Although the grand-average ERP pattern from Kim & Osterhout (2005) indicates that most subjects should pursue option 2, we nevertheless suspected, based on previous work, that such across-subject averages might actually be composed of individuals who differ in which analysis they end up pursuing.

The idea that situations of "semantic attraction" represent a tipping point between syntactically- and semantically-driven analysis was explicitly investigated by Kim & Sikos (2011). They presented subjects with three conditions, reproduced below.

(KS1) The hearty meal was *devoured...* *Original Control Condition*

(KS2) The hearty meal was *devouring...* *Original Semantic Attraction Condition / Single Edit Repair*

(KS3) The hearty meal would *devour*... *Multiple Edit Repair*

The authors reasoned that if semantic P600 effects are elicited in situations in which semantic processing "wins", then it should be possible to strengthen syntactic cues to a point at which they tip the balance back in their favor. Thus the inclusion of the "multiple edit repair" condition above, in which a reanalysis of the sentence leading to a semantically plausible interpretation would involve changing at least two aspects of the sentence: the inflection on the verb from "devour" to "devoured" and the change of the modal verb "would" into the auxiliary verb "was". On the other hand, sentences in the original semantic attraction ("single-edit repair") condition would require only a change of inflection, making it easier for semantic processing to dictate re-analysis of the sentence. The authors predicted that the "stronger" syntactic cues contained by sentences like (KS3) should lead to an N400 effect in response to the anomalous verb, since the structure of the sentence would dictate that "meal" is indeed the agent of the verb. Consistent with predictions, a P600 effect was replicated in the single-edit repair condition, while sentences in the multiple-edit repair condition elicited an N400.

The results of Kim & Sikos (2011) seem to indicate that semantic P600 effects are to an extent "fragile", insofar as subtly strengthening the syntactic cues of a sentence can nearly eliminate them from a grand-average ERP pattern. We therefore hypothesized that the weight which the comprehension system places on syntactic versus semantic cues in the computation of meaning might vary to some extent across individuals. Indeed, there has been previous work showing qualitative differences between individuals in their tendency (or ability) to incorporate relatively salient semantic features, such as animancy, into their comprehension strategies. For instance, Nakano, Saron & Swaab (2010) administered the reading span task (Daneman & Carpenter 1980), a measure of Verbal Working Memory, to a group of native English speakers and recorded ERPs while they listened to three types of sentences, reproduced below.

(N1) The dog is biting the mailman. *Control Condition*

(N2) The poet is biting the mailman. *World-knowledge Violation Condition*

(N3) The box is biting the mailman. *Animacy Violation Condition*

ERPs were analyzed for all of the underlined words. When subjects were split into high- and low-span groups based on their performance on the reading span task, and their ERPs analyzed, it was found that high-span subjects displayed a frontal negativity for initial nouns in the animacy violation condition, followed by a P600-like effect elicited by the verb (again, only for sentence containing animacy violations). Low-span subjects, on the other hand, did not respond differentially to the initial noun in the animacy violation condition, and subsequently showed an N400 in response to the verb. The authors interpreted this as suggesting that only high-span readers could take into account the animacy of the initial noun, and therefore guided their interpretive processes towards expecting a passive verb. Thus, when faced with an *active* verb, they displayed a P600 effect, indicating that the inflection may have caused structural processing difficulties. Since low span subjects were unable (or unwilling), according to Nakano et al (2010), to incorporate animacy information into their comprehension process, they displayed an N400 effect, indicating that the verb was semantically, instead of structurally, difficult to integrate.

If there are individual differences across readers in the strategies used to process sentences as simple as (N1) - (N3), then we might reason that cognitive differences between individuals might also manifest in differing responses to situations involving "semantic attraction". For instance, Nakano et al's (2010) results would seem to predict that high-span readers, able to maintain the animacy information conveyed by the initial noun throughout their processing of the sentence, should be more likely to exhibit P600 activity to *both* of Kim & Osterhout's (2005) experimental conditions, since in all

cases the inanimate noun would signal likelihood of a passive structure. Low span subjects, on the other hand, would be predicted to exhibit *N400* activity in both conditions. However, we might also consider the possibility that sentences in the semantic attraction condition could allow even low-span subjects to consider a passive interpretation, since the semantic relationship between the subject noun and verb might cue the comprehension system to re-analyze the sentence.

To test these ideas, we employed stimuli very similar to those used by Kim & Osterhout (2005), and examined whether there was indeed individual differences with respect to which effect (N400 or P600) sentences in both experimental conditions of the study elicited. Furthermore, to address the question of what cognitive factors might influence an individual's ERPs for these types of sentences, we administered a battery of cognitive tasks (including the Daneman & Carpenter 1980 reading span task employed by Nakano et al. 2010) designed to identify resources or processes that might influence what types of information an individual is able incorporate into their comprehension strategy.

Although the exploration of individual differences in the generation of language-related ERPs is the principal focus of the current study, we took the opportunity to present subjects with another set of stimuli (within the same stimulus lists) in order to explore a different aspect of language comprehension. These stimuli were designed to examine the interaction of semantic processes with early stages of visual word-form recognition, and is an elaboration of work by Kim & Lai (2011), which presented subjects with sentences from the following four conditions.

(KL1) She measured the flour so she could bake a <u>cake</u> …	*Control Condition*
(KL2) She measured the flour so she could bake a <u>ceke</u> …	*Supported Pseudoword Condition*
(KL3) She measured the flour so she could bake a <u>tont</u> …	*No-Support Pseudoword Condition*
(KL4) She measured the flour so she could bake a <u>srdt</u> …	*Nonword Condtion*

When the authors examined the early, occipital-area ERP components – the P100 and N170 – that were elicited by their manipulations to the underlined critical words above, they found that only the Supported Pseudoword Condition modulated the P130 component relative to the control condition, while the two other experimental conditions modulated the N170. This was taken as evidence for early, rapid interactions between semantic areas and visual areas that occur when input resembles, but does not totally match, expectations generated by the preceding context. The Authors conjectured that the modulation of the P100 in the supported pseudo-word condition occurred when bottom-up input partially activated an already primed, higher-level semantic representation of "cake", which in turn sent top-down feedback to early visual wordform areas, ultimately leading to an early-stage conflict between visual representations for "cake" and those activated by the actual input "ceke". Since this result was unexpected given some non-interactive models of visual word recognition, we included a modified version of Kim & Lai's (2011) materials in our experiment to both replicate and further explore interactions between visual and semantic processing. Our manipulations involved changing the pseudowords in the supported and no-support conditions to misspellings of the control words, such that instead of "ceke" and "tont" we used "ckae" and "tnet" (where "tnet" is a variation of "tent", which occurs as a supported misspelling for another stimulus item), respectively. We hypothesized that keeping the constituent letters of the control word the same while manipulating their order might modulate the amplitude or latency of the P100 effect in the supported pseudoword condition, since it would isolate the conflict to a stage of visual processing concerned strictly with the configuration of visual letter-forms, given that the same visual letter representations would be activated in both cases. We also included a condition that contained a misspelling of a word that was semantically related to words in the preceding context, but was nevertheless not appropriate given the meaning of the sentence as a whole (e.g. "The campers finally found a great place to light a *tnet* on fire"). We believed that

observing a P130 effect in this condition would indicate that anticipatory activation of visual word-forms is "broad", in the sense that perception of one word will automatically prime later recognition of closely related ones, regardless of whether they would be contextually appropriate. On the other hand, if a P130 is observed for the supported condition but not this "mid-supported" condition, it would be fairly strong evidence that anticipatory priming of visual word-forms is dependent on the preceding context as a whole, and that only word-forms which continue the narrative of a sentence in a plausible way become primed for recognition.

Although all the data that we describe below were obtained from the same set of subjects, we have partitioned the remainder of the paper according to the three principle approaches we took in analyzing it. First, we give a description of the Grand-Average ERP pattern obtained by presenting subjects with stimuli similar to those employed by Kim & Osterhout (2005). Then we describe our approach for characterizing individual differences in responses to these stimuli, as well as the independent cognitive measures that we believe partially accounts for this variability. Finally, we describe the results from the portion of our study concerned with how varying levels of contextual constraint affect the process of visual word recognition. We treat each of these three approaches as separate experiments, and discuss the results of each in turn below.

EXPERIMENT 1: GRAND-AVERAGE ERPS ELICITED BY SEMANTIC ATTRACTION STIMULI

II. METHODS

Participants

EEG Data were collected from 50 right-handed native English speakers with normal or corrected-to-normal vision, ranging in age from 18 to 27 years old. Of these, five subjects were excluded from individual differences analyses for not completing the behavioral portion of the

experiment. To minimize the influence of noise on individual subject ERPs, we also set a criterion that at least 34 trials be retained, after artifact rejection and correction, for each of the 7 major conditions in the EEG portion of the experiment, which excluded five further subjects from individual differences analyses. Data for a total of 40 participants were therefore retained for individual differences analyses, reported in experiment 2. Participants were compensated with either $35 or course credit. All participants gave written consent. All procedures described below were approved by the University of Colorado Human Resources Council and Institutional Review Board.

Materials

We employed the original 96 stimuli from Kim & Osterhout (2005) as well as 24 additional items created to increase the reliability of single-subject ERP averages. Each item consisted of three versions, represented by sentences (SA1)-(SA3) below.

(SA1) The hearty meal was devoured by the boys *Control*

(SA2) The hearty meal was devouring by the boys *Attraction Condition*

(SA3) The dusty tabletop was devouring by the boys *No Attraction Condition*

Each item had its three versions distributed across different lists (see "creation of stimulus lists", below) in such a way that no subject saw more than one version of each item. As discussed above, control stimuli were passive sentences in which the subject was a highly plausible theme for the verb. Sentences in the attraction condition were created by changing the inflection of the main verb in the control sentences from passive (-ed) to active (-ing). Sentences in the no attraction condition were syntactically identical to the attraction sentences, but contained subjects that were not plausible themes

for the verb. All items contained subjects that were inanimate nouns.

Creation of stimulus lists (Common to all Experiments)

Four main stimulus lists were created by distributing equal amounts of visual word form (see experiment 3) and semantic attraction items to each list, with equal amounts of each item per condition (40 per semantic attraction condition, and 45 per visual wordform condition). Since the majority (approximately 72%) of sentences were in some way anomalous, 130 well-formed filler sentences of varying length and syntactic complexity were added to each list to make the ratio of well-formed to anomalous sentences even. Each main list (consisting of 430 sentences in total) were pseudo-randomized in such a way that no three sentences from the same condition occurred in sequence. Four additional stimulus lists were constructed by reversing the presentation order of the original four lists.

EEG Recording Procedure (Common to all Experiments)

All subjects completed the EEG portion of the study before completing the behavioral tasks at a later date. Subjects sat quietly in a dark a room and were told not to blink or move while they viewed sentences on an LCD monitor approximately three feet away. Subjects were told to monitor for unusual sentences, and were given examples of normal sentences, sentences containing misspellings, grammatical anomalies and semantic anomalies. Sentences were displayed one word at a time in white text on a black background (Rapid Serial Visual Presentation (RSVP)). Words were displayed for a duration of 380 ms with a blank inter-stimulus interval of 100 ms. Following each sentence a prompt was displayed asking subjects to judge whether or not the sentence was normal (that is, did not contain any misspellings, grammatical or semantic anomalies). Subjects made a button press to indicate their

response. Order of button presses (left or right handed for normal or not normal) was counterbalanced across subjects. Subjects saw one of eight possible stimulus lists and were given breaks after every 72 sentences. Following completion of the experiment subjects were debriefed on the study's purpose.

Recording and Processing of EEG Data (Common to All Experiments)

Continuous EEG was sampled at a rate of 1000 Hz with a 66 electrode cap (Compumedics Inc), and amplified and digitized with a SynAmps2 amplifier. Recorded data was subsequently down-sampled to 200 Hz and band-pass filtered between .1 and 50 Hz. Changes in EEG voltage exceeding +/- 100 microvolts were rejected from further analysis. Correction of eyeblink artifacts was completed using a subject-specific regression-based algorithm (Semlitsch, Anderer, Schuster & Presslich, 1986).

Subsequent to recording, continuous EEG data were cut into epochs of 1000 ms time-locked to the onset of the critical stimuli. Baseline correction was done using a 100 ms pre-stimulus baseline period. Epoched data were averaged within experimental conditions (*control, attraction* and *no attraction*) for each subject and re-referenced to the average of the mastoid channnels. Resulting subject averages were then used to complete across-subject grand averages. Resulting grand-average ERPs for a centrally located region of interest (ROI) are displayed in Figure 1.

III RESULTS

Visual inspection of grand-average ERPs elicited by all three conditions revealed that we were largely able to replicate the qualitative ERP patterns from Kim & Osterhout (2005), although we did observe a couple of surprising results. First, although we observed the predicted P600 effect in the attraction condition, the positivity was both smaller (relative to control), and at longer latency than has

been previously reported. Secondly, *both* conditions elicited what appears to be a clear N400, with sentences in the 'no attraction' condition showing a comparatively larger effect. While consistent with our predictions for the no attraction condition, the N400 observed for the attraction sentences was somewhat surprising given the original results of Kim & Osterhout (2005) as well as Kim & Sikos (2011).

Based on visual inspection of the grand-average waveforms in figure 1, we chose time windows of 280 to 550 ms and 550 to 900 ms, corresponding to the latencies of the N400 and P600 effects, respectively, to statistically evaluate the foregoing effects. We decided to employ relatively broad time-windows to allow for individual variation in the onset of N400 and P600 components, which were analyzed on an individual basis in experiment 2. Voltages in these time windows were averaged for electrodes in a central ROI (Channels PZ, CP1, CZ, CP2 and CPZ) located on the midline of the scalp (see figure 2). Results were submitted to a repeated measures analysis of variance with factors latency (two levels: N400 and P600 time windows) and condition (three levels: control, attraction and no attraction). Subject served as the random factor. Results from the analysis of variance, as well as pair-wise comparisons between conditions, are reported in table 1 and table 2, respectively.

In the N400 time-window, results indicated that the no attraction items elicited a significant N400 effect relative to control sentences [$F(1,39) = 18.711$, $p < .001$]. A pair-wise comparison between attraction and control conditions also indicated the presence of a smaller N400 effect [$F(1,39) = 7.06$, $p = .012$].

In the P600 time-window, results confirmed the presence of a significant P600 effect for the attraction condition relative to control [$F(1,39) = 7.063$, $p = .011$]. Comparison of no attraction items to control items did not, however, reveal a significant positivity [$F(1,39) = .734$, $p = .40$].

Discussion

We employed the original stimuli (in addition to ten extra items per condition) from Kim & Osterhout (2005) with the anticipation that a replication of their grand-average (that is, a P600 in the semantic attraction condition and an N400 in the no attraction condition) pattern would provide us an opportunity to explore possible individual differences in the generation of these effects. The resulting grand-average largely confirmed predictions, with some minor deviations. Because this is the first replication our lab has attempted of the no attraction condition, we expected any deviation from our predictions to be caused by these stimuli; instead, it was the semantic attraction condition that yielded the two most surprising aspects of these results, namely (1) a small but significant N400 effect and (2) a smaller – and later – than expected P600 effect.

Though the sources of these unexpected effects are difficult to determine with certainty, the major difference between the current study and past studies in which we have employed these items is the addition of the stimuli that constitute the visual wordform portion of the design (described in experiment 3). Sentences in three of the four conditions from the VWF design contained misspellings, which are known to elicit the strong P600-like effects that we ultimately observed in our own data. Previous work has also established (e.g. Hahn & Friederici 2001) that instructing subjects to detect anomalies tends to enhance P600 effect sizes to known triggers, and in some cases the effect can be shown to disappear completely if subjects are instead told to read passively or ignore syntactic errors. Since our task required subjects to judge the well-formedness of every sentence, it may be that (on average) subjects were more sensitive to the spelling anomalies, which outnumbered the semantic attraction items more than three to one. Indeed, it can be argued that the misspellings are more readily detectable, since subjects must pay enough attention to the morphology of the verbs in the semantic attraction condition in order to judge them as anomalous. We therefore suggest that both the increased number and detectability of the misspellings may have implicitly drawn attention away from the

semantic attraction items and therefore contributed to the unexpectedly small P600 effect that we report here.

We also observed a small, but significant, N400 effect in the attraction condition, which is also somewhat inconsistent with past research that has employed these stimuli. The effect is surprising given the fact that items in this condition were designed to contain a strong semantic relationship between the verb and the subject noun, and thus should, notwithstanding the unexpected inflection on the verb, attenuate any ERP effects due to violations of semantic expectations. We therefore hypothesized that this effect may be due to a minority of individuals that showed an N400 in the semantic attraction condition, without showing any indication of the predicted P600. Indeed, of the eight subjects that did show an N400 effect in the attraction condition, only three of them showed a following positivity, indicating that the majority of this effect was driven by a group of five individuals with an average effect size in the N400 window of just over three microvolts.

The results still provided us, however, with an opportunity to investigate possible individual differences in the generation of the N400 and P600 effects we observed. Indeed, it may have been that our (somewhat atypical) grand-average pattern may actually be due to variability in the composition of our sample, which may have included subjects whose individual brain responses differ from the "standard" N400/P600 dichotomy that previous studies have reported. If this is the case, then we may actually be in luck, since a healthy variability among subjects is crucial for effective individual differences research. In the following sections we will outline our method for examining how the grand-average pattern in the semantic attraction portion of the study might be analyzed in terms of contributions from different types of subjects, and what cognitive differences among these subjects might explain any individual departures from the brain patterns we observed in the grand-average.

EXPERIMENT 2: INDIVIDUAL SUBJECT ERPs ELICITED BY SEMANTIC ATTRACTION

IV. METHODS

Participants, stimuli, EEG recording procedures and post-processing were identical to those of experiment 1.

Procedure - Collection of Behavioral Measures

After completing the EEG portion of the study, subjects were scheduled for a second session to complete the battery of behavioral tasks, during which they performed the following seven tasks, in the order shown:

(1) Vocabulary Test

(2) Magazine Questionnaire

(3) Author Questionnaire

(4) Spatial Two-Back

(5) Keep-Track Task

(6) Spatial Span Task

(7) Reading Span Task

In the vocabulary test (Educational Testing Service, 1964), subjects were asked to identify, for 36 words, which of five choices best represented the meaning of the word in question. Subjects were told not to guess if they did not know the answer. The dependent measure was the proportion of questions answered correctly.

For the Magazine and Author Questionnaires, subjects were provided a list of 80 magazine titles or Author names, half of which were real titles or names and half of which were not. Subjects were then asked to mark which titles and names they recognized. The dependent measure was the proportion

of items identified correctly. The questionnaires were designed to be a quick measure of language exposure.

In the spatial two-back task, which was designed to test the ability of subjects to flexibly update spatial representations in working memory, participants viewed an array of white boxes on a computer screen. Every two seconds a box was turned black, and subjects indicated with a button press whether or not the current box was the same box that was turned black two trials back. The dependent measure was the proportion of targets correctly identified.

In the Keep-track task, designed to test the ability of subjects to flexibly update verbal representations in working memory, participants viewed a computer screen as words from six different categories (animals, colors, metals, relatives, countries and distances) appeared one at a time. Subjects were told to keep track of between two and five categories for each trial, and to report back the last item from each category at the end of each trial. The dependent measure was the proportion of words correctly recalled.

In the spatial span task, designed to evaluate an individual's spatial WM capacity, participants viewed alternating sequences of rotated letters and arrows on the center of a computer screen. Subjects were required to judge whether the rotated letters were normal or mirror-imaged, and at the end of each trial to recall the direction and serial order of all arrows (between two and five) from that trial. The dependent measure was the proportion of arrows correctly recalled.

In the reading span task, designed to evaluate an individual's verbal WM capacity, participants viewed alternating sequences of English sentences and single words on the center of a computer screen. Subjects were required to judge the truthfulness of each sentence, and at the end of each trial to recall the serial order of each singly presented word (between two and five). The dependent measure was the proportion of words correctly recalled.

The magazine and Author questionnaires were added after approximately 13 subjects had

already participated in the experiment without them, so the reader must keep in mind that correlations of these measures with ERP data may not be as reliable as those computed with the other five tasks.

V. RESULTS

Behavioral Measures

Table 3 shows summary statistics for the behavioral measures. Since the dependent measure for all behavioral tasks was the proportion of items correct, we applied the arcsine transformation (following Judd & McClelland 1989) to ensure a normal distribution for each of our individual differences measures, as well as to help guard against floor and ceiling effects. However, Anscombe-Glynn tests for kurtosis revealed that, even after transformation, distributions for the Author questionnaires and spatial 2-back tests remained marginally kurtotic (p-values .081 and .084, respectively), indicating the possibility of floor and ceiling effects, respectively, for these tasks.

Table 4 displays the correlation matrix for these behavioral tasks.

Individual Subject ERPs

Our primary method for characterizing individual differences in sentence processing is described below. The approach was two-pronged: first, we developed a method for characterizing an individual's brain response to linguistic stimuli based on the relationship between their N400 and P600 effect sizes within each condition. In the second step of our analysis we related these individual brain measures to performance on a battery of cognitive tasks in an attempt to examine how more domain-general cognitive abilities contribute to and shape an individual's sentence processing strategies and capabilities. This two-step approach allows us to investigate both to what extent individual ERP

signatures deviate from (or conform to) the grand-average N400 and P600 patterns that are typically observed in psycholinguistic research, as well as characterize to what extent these individual patterns interact with other cognitive processes that are not strictly bound to language processing.

The N400-P600 Continuum: A Derived Dependent Measure

We suspected that some aspects of our grand-average ERP pattern may have been an amalgam of different classes of individual ERP patterns, some of which might deviate considerably from the grand-average. Specifically, in terms of the grand-average for these data, we felt that the small N400 effect in the attraction condition, and what appeared to be a smaller, late positivity in the no attraction condition may be due to two minority groups of subjects that deviated from the grand-average pattern. This was driven principally by the observation that, while individual subjects largely conformed to our predictions in the two most critical time-windows (with 83% of our subjects showing an N400 in the no attraction condition, as well as 80% showing a P600 in the attraction condition), there was considerable variability in the P600 time-window for the no attraction condition (approximately 48% of subjects showed a positivity) as well as in the N400 window in the attraction condition (approximately 41% showed a positivity). This led us to ask whether the mildly bi-phasic pattern of effects found in the grand-average (that is, a small N400 followed by a large P600 in the attraction condition, and a larger N400 followed by a statistically insignificant P600 in the no attraction condition) may have been produced by averaging across different "classes" of subjects that tended to show, within a condition, one effect but not the other. This amounted to asking whether or not, within subjects, effect sizes in one time window were *anticorrelated* with effect sizes in the other window.

To establish statistically whether N400 and P600 effects were anticorrelated within each condition, we computed two linear regressions predicting P600 effect sizes from N400 effect sizes. For

calculation of individual N400 effects we used the same time-window employed in the grand average (280-550 ms), but for the P600 effects we chose a later time window of 680 to 990 ms, creating a 'gap' between the two time-windows of 130 ms. We decided to employ a "buffer" between time windows to help guard against the possibility that the regression analyses would reflect effects that overlap between the two time-windows. Though we cannot fully control for this possibility across all participants, we believe that employing *effect sizes* (control voltage subtracted from anomalous voltage) as our unit of analysis, rather than raw voltage deflection from baseline, helps address the possibility that our time-windows might be partially non-independent.

In both conditions, N400 effect size turned out to be a significant predictor of P600 effect size [attraction condition: $b = .441$, R-squared = .15; no attraction condition: $b = .441$, R-squared = .18] (see table 5 and figures 3 and 4 for a graphical representation), indicating that an individual subject's response to both of our manipulations in the N400 time window seemed to partially determine their ERP response starting 130 ms later. We also computed two similar regressions *across* conditions (see table 5), such that N400 effect size in the no attraction condition served as a predictor for P600 effect size in the attraction condition, and vice-versa. In these cases, N400 effects ceased being a significant predictor of P600 effects, indicating that the relationship is dependent on our experimental manipulations.

To help minimize the influence of outliers on these (as well as all subsequent) analyses, we computed values for Cook's *D,* which is a measure of how much each observation influences the parameter estimates of a model. Observations whose Cook's *D* values greatly exceeded the mean for each model were treated as outliers and removed from the analysis.

The slopes from our regression analyses indicated that the size of a subject's P600 effect is indeed largely *anticorrelated* with how much their ERP deflected from control in the N400 time-window (and vice-versa): subjects who had a tendency towards large N400 effects tended to show

smaller, or non-existent P600 effects. This implies a sort of “trade-off” between P600 and N400 effects within subjects, such that few display large effect sizes in both time-windows. To ensure that subjects on both extremes of the spectrum were generating ERPs that qualitatively resembled traditional N400 and P600 components, we computed two sets of across-subject ERPs in each of the two experimental conditions for the six subjects who displayed the largest N400 and P600 effects.. These results (see figures 5-8) indicated that in each case the waveforms appeared to conform to the general shape of the ERPs from the grand-average.

In all, the results suggest the existence of an N400-P600 'continuum', in which subjects respond to our anomalous stimuli more-or-less exclusively with one effect or the other. Since we believe this to be in line with our understanding of what each component reflects in terms of processing "decisions", we were interested in exploring what factors might contribute to a given subject's place on this 'continuum' of ERP effects, and, presumably, comprehension strategies.

Quantifying Position on the N400-P600 Continuum

Since this segregation of subjects by component effect size was purely data-driven, we wanted to ask whether we could relate this N400-P600 “trade-off” to other more independent measures, specifically, the performance of these subjects on our behavioral tasks. As a result we developed a metric which we believe can reliably quantify each subject's position on the N400-P600 “continuum” with one value, making it amenable to regression analyses in which we could employ our behavioral measures as predictor variables. Figure 9 displays the creation of this metric graphically. In essence, we took the regression lines resulting from our analyses that related P600 and N400 effect sizes, and used that as a single dimension in which to express a subject's position on the N400-P600 continuum for each experimental condition. Individual subject values on this continuum were calculated by projecting

their original coordinates in N400-P600 space orthogonally onto the regression line. Figure 10 shows these transformed values graphically. Note that these values are different from what would result if we used the coordinates of the *predicted* values of P600 size given N400 size. Orthogonal projections were chosen because of the ambiguity introduced by the minority of individuals who have positive effect sizes in one condition but negative effect sizes in the other. To illustrate this problem, consider the case of subject 17, circled in figure 11. This individual has relatively large but *opposing* effect sizes, making it difficult to characterize him or her as either 'more N400-like' or 'more P600-like'. If we compare the position of this subject on the "continuum" after orthogonal projection versus where they would be if we used predicted values from the regression analysis, we see that subject 17 ends up more towards the negative side of the continuum when predicted values are used. Since this subject is poorly characterized as either just positive or just negative, we should instead hope to see their value lie somewhere more towards the middle of the distribution, which is what orthogonal projection appears to achieve. Thus, while our scheme cannot directly account for *why* these people depart from the N400-P600 dichotomy, it at least relegates them to the middle of the distribution, where they will have less influence on further regression analyses.

To quantify each subject's position on the N400-P600 continuum, we calculated the Euclidean distance between each subject and the y-intercept of the regression line (see figure 10), with subjects lying to the left of this point given negative values. Using this 'continuum metric' we investigated whether a subject's N400 or P600 'tendency' was related in any way to their performance on our behavioral tasks. Taking the continuum measures as our dependent variable, we initially computed a series of simple regression analyses to determine which, if any, of our behavioral tasks predicted the subjects' average responses to our anomalous conditions.

Behavioral Measures that Predict Positions on the N400-P600 Continuum

Results of regression analyses relating our behavioral measures to individual positions on the N400-P600 continuum are summarized in table 6. A significant relationship was found between our N400-P600 continuum measure for the no attraction condition and scores on our verbal WM updating task [b = 10.379, p = .003, R-squared = .227], indicating that subjects with lower scores tended to show N400 effects, while higher scoring subjects were more likely to display P600s. Figures 12 and 13 display ERPs from the no attraction condition for the six highest and lowest scoring subjects on the keep-track task. Additionally, a marginally significant relationship was found between the continuum measures (again for the no attraction condition) and our measure of spatial working memory span [b = -3.985, p = .058, R-squared = .108], indicating that higher scoring subjects were more likely to show an N400 effect. Graphical representations of these relationships can be found in figures 14 and 15, respectively. Importantly, we found that none of our behavioral tasks predicted the N400-P600 continuum measures for the semantic attraction condition. We hypothesized that this might be due the comparatively smaller amount of variance in P600 window activity for this condition: nearly all subjects showed a positivity in response to our semantic attraction stimuli. Furthermore, we observed no correlations between our ERPs and the reading span task, which both Nakano et al (2010) and Bornkessel, Fiebach & Friederici (2004) previously found to be related to individual differences in N400 and P600 effects.

VI. EXPERIMENTS 1 & 2: DISCUSSION

Implications of Individual Variability in ERPs for the Interpretation of Grand-Averages

The fact that grand-average ERP patterns, taken to be indicative of any given subject's

processing strategy, may potentially be combinations of *different* responses from different types of individuals could have implications for how we interpret functional changes in the language system that we ascribe to the modulation of a given ERP component. Crucially, this fact would not *necessarily* change the functional interpretation of any one ERP component, but it *does* mean that any functional interpretations that we make about ERP effects observed within a grand-average must be tempered in proportion to the consistency that we observe that component on a subject-by-subject basis. Thus, in the case of our data, the interpretation of the grand-average N400-P600 pattern elicited by no attraction stimuli needs to be made in light of the fact that there are underlying processing differences occurring *between,* instead of *within,* subjects. This means, as we will argue in detail below, that subjects showing an P600-like response to our no-attraction anomalies may be processing these stimuli in ways that differ significantly from their N400-like peers.

To be clear, the fact that our grand-averages can, in some cases, be characterized as amalgams of different types of subjects does not lead to the firm conclusion that there are, as a general rule, N400-like people and P600-people. Not only do other datasets with similar stimuli need to be examined to determine whether this individual variability is contributing significantly to other to other grand-average patterns, but we can observe directly in our own data that the tendency of an individual subject to "show" a particular ERP component in response to no-attraction items (for example) is a poor predictor of their responses to items from other experimental conditions. Thus the alternation between N400 and P600 within subjects is a complicated matter, since each subject's ERP 'profile' exists on a continuum between N400 and P600 responses, with the conditions necessary to elicit either component varying considerably between different types of individuals. Below we attempt to identify, based on findings from our behavioral measures, what cognitive qualities help determine an individual subject's response to the types of anomalies we included in our research.

Individual differences that predict the N400-P600 trade-off to Stimuli in the No Attraction Condition

Following the results of Nakano & Swaab (2010) (as well as others) we expected the best predictor of individual P600/N400 amplitude to be the reading span measure, which has traditionally been one of the primary tools used to relate WM capacity to reading performance. Instead, two other individual differences variables – WM updating performance and Spatial Working memory - were found to be better predictors of N400 -P600 continuum measures that were obtained from the no attraction stimuli. Furthermore, *none* of our behavioral measures predicted responses to our semantic attraction stimuli. This may have been due to a surprising lack of variance across subjects in the P600 window for the semantic attraction items, indicating that the effect is more reliable than we had anticipated. In turn, this would seem to indicate that most subjects, independent of any cognitive differences that existed between them, were able to consider the semantically-driven interpretation of "the hearty meal was devouring...". This would also seem to imply that, at least based on Kim & Osterhout's (2005) (as well as Kim & Sikos's 2011) interpretation of semantic P600 effects, that the majority of subjects had "enough" cognitive resources available to re-analyze the syntactic cues of the sentence in order for them to properly reflect the semantically plausible interpretation. Since WM span (typically as measured by the Daneman & Carpenter 1980 reading span task) has been previously implicated as a critical measure of the resources necessary to perform online re-analysis of structural representations (especially in garden-path situations, c.f. Just & Carpenter 1992 for a review), the present results seem to suggest that these particular re-analyses were simple enough that even low-span readers could perform them. Importantly, this interpretation *also* predicts that processing of our no-attraction items should place equal demand on WM resources, since they did not differ *syntactically* from our semantic attraction stimuli. However, the fact that our measure of WM updating was able to predict whether a subject showed an N400 or a P600 to these stimuli (but *not* the attraction items)

seems to contradict this prediction. Because these two conditions differ only in their semantic qualities, this would seem to indicate that WM resources are crucial not only for the processing of syntactic information, but semantic information as well. Specifically, as will be outlined in detail below, encountering a semantically unexpected word (such as the target verbs in our no attraction items) might place an unusually high demand on the comprehension system when it attempts to update its discourse representation with semantically incongruous material.

Working memory updating, or the ability to flexibly load and delete representations from working memory, would seem to be, at least in a priori terms, a crucial component to the incremental, on-line interpretation of language. Each new word that is added to an ongoing discourse must not only be retrieved from long-term memory, but also integrated into both the structural *and* semantic representation of the current sentence or phrase. Garden-path situations (e.g., Trueswell & Tanenhaus 1994) are perhaps the most striking example of a situation in which the comprehender must radically update their representation of the unfolding sentence, but in principle the addition of any word that is not firmly predicted by the comprehender must elicit some change to the current semantic and structural representation of the discourse. Since both semantic and structural aspects of an unfolding sentence can in principle be updated to reflect the contribution of each incoming word, it is reasonable to assume that updates involving *both* types of information may be more taxing to the language system than those which involve only one or the other. The anomalies that we included in our stimuli, we believe, presented subjects with a situations in which the required change to the discourse representation was either mainly structural, or both structural *and* semantic. Since all sentences in this portion of the design began with an inanimate noun, the critical verb should be expected to have an -ed inflection, based on the fact that sentences containing inanimate subjects are more likely to be passive (c.f. Master 1991, Dewart 1979). Thus, critical verbs in the attraction condition represent situations in which the target is semantically predictable, but the inflection or structural information contradicts the

expectation of passive structure. Since, in this condition, the semantic content of the critical verb was relatively predictable, the principal stress on the comprehension system will involve incorporating the unexpected syntactic information indicated by the -ing verb inflection. By contrast, critical verbs in the *no* attraction condition represent situations in which both the semantic *and* structural information conveyed are unexpected, requiring the system to update *both* types of information simultaneously, which is presumably a more difficult operation.

The forgoing interpretation of our attraction and no attraction conditions suggests that subjects with better performance on updating tasks might show different brain responses when presented with words (such as the critical verbs in the no attraction condition) that place a high demand on the language system to update its representations of the incoming discourse. In fact, the tendency of subjects to show either a P600 or N400 (as indicated by N400-P600 continuum measure) to critical items in the no attraction condition was significantly predicted by their performance on the keep-track updating task, with better performing subjects more likely to exhibit P600-like activity. The fact that low-updating subjects were more likely to show pronounced N400 effects to these types of sentences suggest that they expended most of their resources on incorporating the unexpected semantic information into the discourse, while (perhaps) failing to fully update their structural representations, or consider a syntactically unlicensed passive analysis. High scoring subjects, on the other hand, may have been able to address both the unexpected semantic *and* structural information conveyed by the verb, since they were more likely to show P600-like activity, which has been previously associated with situations that involve demands on structural processing.

The foregoing account also predicts that critical items in the attraction condition should place *less* demand on the system to update its representations, since their semantic qualities were highly predictable given the context, and therefore likely already (to some extent) integrated into the comprehender's discourse representation. Indeed, when we calculated N400-P600 continuum measures

by regressing P600 size on N400 size in the attraction condition only, the keep-track task failed to predict the N400-P600 trade-off. Together, the data seem to suggest that only our no attraction condition provided subjects with a situation difficult enough to reveal how individual differences in working memory performance might affect processing of linguistically problematic situations.

Two important assumptions underlie our interpretation of the relationship between working memory updating and individual N400/P600 effects. First, the account assumes that online interpretation is highly predictive, and that the ease of integrating into the discourse both the semantic and structural information contained by a word depends largely on how well it fits with the comprehender's expectations. We believe this to be a reasonable assumption, given the wealth of evidence supporting the idea of semantic and syntactic prediction (see e.g., Altmann & Kamide 2003 Kim &Lai 2011). Perhaps more controversially, our story also assumes that updating of semantic information may require the *same* set of resources as structural updating, since we are claiming that low updaters exhaust their abilities on semantic processes when faced with the critical verb in the no attraction condition. Since there exists an extensive literature (c.f. Caplan & Waters 1999) suggesting that structural processing is the main consumer of working memory resources, we must leave validation of this assumption to further research.

It is also worth noting that our results are also partially consistent with that of Nakano & Swaab (2010), which showed that subjects scoring well on measure of verbal WM span showed a P600 in response to animacy violations. Although we used the same span task and failed to uncover a relationship between performance and ERPs in either of our experimental conditions, our WM updating task appears to have segregated ERPs in our no attraction condition in a similar way, with higher-scoring subjects showing a P600 to sentences such as “the dusty tabletops were devouring”, which are also clear animacy violations. Interestingly, however, the majority of our subjects, also showed the predicted P600 effect in response to the semantic attraction stimuli, which are also – at least on the face

of it – animacy violations. Since our WM updating measure failed to predict ERPs in this condition, we cannot firmly conclude that our P600 effects were due entirely to violations of animacy, though it is conceivable that something about our semantic attraction manipulation allowed more subjects (not just high-updaters) to respond to such violations. One possibility, consistent with Kim & Osterhout's (2005) account of semantic attraction, is that the semantic relationship between the subject and verb provided another cue (above and beyond either animacy or verb inflection) which allowed subjects of all WM abilites to overturn the syntactically-signaled interpretation of the sentence and consider a semantically-driven interpretation. In any case, further research will be needed to better characterize the relationship between individual WM abilities and P600 effect size.

Besides our working memory updating measure, it was also found that scores on our measure of *spatial* working memory span, the so-called spatial span task, also predicted the N400-P600 trade-off that we observed in the centrally-located ROI. The spatial span task is a so-called "complex" working memory task and therefore is likely to engage several cognitive functions that underlie the processing and manipulation of visuo-spatial information. The task involves viewing an alternating sequence of rotated letters and and arrows, with subjects being tasked to identify whether the letters are normal or mirror-imaged, and to recall (in order) the directions of the arrows at the end of each trial. Thus the task involves both the manipulation (judgment of the rotated letters) and maintenance (memorization of the arrow directions) of visual information. Since the task was originally included to serve as a comparison to the reading span task (in order to ascertain whether or not the resources tapped by the reading span task were verbal-specific), it is surprising that it can serve as an effective predictor of our language-related ERPs. Since both the keep-track and spatial span scores remain significant in regression analyses when the other is entered as a covariate, the usefulness of the spatial span measure as predictor must go beyond any aspect of the task that is measuring some sort of verbal working memory updating. Furthermore, the spatial span scores predict ERPs in the opposite way from the keep-track data: higher

scoring individuals tend to elicit more N400-like activity and less P600-like activity to all anomalies in the semantic attraction portion of the study. The complex nature of the spatial span task makes it difficult to pinpoint what component of it might demand the use of processes that are also employed during language comprehension, so, given that the effect is replicable, further research that employs more basic spatial tasks as individual differences variables will be needed to illuminate exactly what aspects of this task might be related to language processes.

EXPERIMENT 3 - EFFECTS OF CONTEXTUAL CONSTRAINT ON EARLY STAGES OF VISUAL WORDFORM RECOGNITION

VII. METHODS

Participants and Procedure

The same participants who completed experiment 1 and 2 completed experiment 3 during the same EEG session. EEG recording and processing remained the same for this portion of the study. Stimuli from this experiment were inter-mixed in lists with stimuli from experiment 1, according to the procedures previously described.

Stimuli

Experiment 2 was designed to investigate the sensitivity of early posterior ERP components to manipulations of contextual support. The experiment employed 180 items, a portion of which were

adapted from Kim & Lai (2011). Each item consisted of four versions, represented by sentences (VWF1)-(VWF4) below.

Control Condition:

(VWF1) The backpacker finally found a campsite and set up the tent before dark.

Contextually Supported Misspelling Condition:

(VWF2) The backpacker finally found a campsite and set up the tnet before dark.

Contextually Unsupported Misspelling Condition:

(VWF3) The backpacker finally found a campsite and set up the sfae before dark.

Semantically Supported, Contextually Unsupported Condition:

(VWF4) The campers finally found a great place to light a tnet on fire.

Each item had its four versions distributed across different lists (see "creation of stimulus lists", above) in such a way that no subject saw more than one version of each item. Control sentences were designed to maximize the predictability of the target words (underlined above), all of which were nouns. In the "contextually supported misspelling" condition (VWF2), target words were misspellings created by swapping the middle two letters of the corresponding control target. In the "contextually unsupported misspelling" condition (VWF3), targets were contextually unpredictable misspellings of control words from other items. Targets in the semantically supported, contextually unsupported condition (VWF4) were identical to targets in conditions VWF1 and VWF2, but embedded in contexts that were not predictive of correctly spelled version of the target, but which also contained two words

that were *semantically* related to the correctly spelled version of the target. All target words were either four or five letters in length. Sentential position of the target words varied across items.

VIII. RESULTS

Grand-average ERPs

Grand-average ERPs for the visual word form portion of the study were computed in a similar manner to the methods employed in the semantic attraction portion of the study, but using ROIs in the left and right posterior portions of the electrode montage for analysis of the early components, and a centrally located ROI (identical to the one employed in the semantic attraction portion) for examining possible N400 and P600 effects (figure 16 shows where these ROIs are located in our electrode montage). Additionally, a 200 ms baseline was used to ensure close alignment of early components. Results can be viewed in Figures 17 and 18, respectively. Since this portion of the study concerned early occipital ERP components, time-windows of 80 to 200 ms and 200 to 340 ms were chosen for statistical analyses of the P100 and N170 components, respectively. A planned repeated-measures ANOVA with factors condition (four levels: control, supported, unsupported and mid-support) time window (two levels: 80-200 ms and 200 to 340 ms), and location (two levels: left posterior and right posterior) was performed with average voltage as the dependent variable and subject serving as the random factor. Results indicated a significant effect of condition [$F(3,93) = 5.42$, p=.0018], however pair-wise comparisons between conditions revealed that the three experimental conditions differed significantly only from the control condition and not from eachother, in either time window or ROI. A summary of statistical analyses performed on these early ERPs can be viewed in tables 7 and 8.

For the later ERPs observed in this portion of the study, a repeated-measures ANOVA with

factors condition (same as above) and time window (280 to 550 ms and 550 to 900 ms) was performed. Results can be viewed in table 9. Statistical analyses indicated a significant main effect of condition [$F(3,37) = 17.61$, $p < .001$], with significant pair-wise differences in the N400 time-window between supported and unsupported conditions, as well as supported and mid-supported conditions [$F(1,39) = 41.5$, $p < .001$, and $F(1,39) = 43.92$, $< .001$, respectively]. In the P600 time-window, all six pair-wise comparisons between conditions indicated significant differences [Fs > 14, ps $< .001$, except unsupported versus mid-supported conditions, $F = 5.68$ $p = .023$]

Relationship of ERPs to Behavioral Measures Collected for Experiment 2

In a more exploratory vein, we were also interested in observing whether any of our individual differences variables could serve as predictors of our early posterior ERP components. Since P100 amplitude size did not vary statistically between conditions in the grand-average of our VWF data, we collapsed across all four VWF conditions and used the averaged individual subject ERPs as the dependent variable in same the time window that was employed in analysis of the grand-average data (80 to 200 ms).

In the N170 time window (300-340 ms), since all three experimental conditions differed from control (but not from each other), we employed the difference between the average of our three anomalous conditions and the control condition as our dependent variable in computing regression analyses in which our behavioral measures served as independent variables.

Since these analyses were largely exploratory, we confined our attentions to ERPs recorded from the left posterior ROI, where effects of visual-word form manipulations have traditionally been the strongest.

Results of the regression analyses relating early posterior ERP components to our individual differences variables in the left posterior ROI can be seen in table 10. Interestingly, there appears to be a moderate correlation between individual P100 amplitudes (collapsed across conditions) and vocabulary size as measured by our multiple choice vocabulary task [$b = 4.48$ $p = .034$, R-squared = .15]. Figure 19 shows a plot of this relationship, indicating that individuals with more positive-going P100 components tended to score higher on our vocabulary measure.

We also observed a marginally significant relationship between N170 effect sizes (average of anomalous conditions minus control) and the reading span task [$b = 5.13$, $p = .058$, R-squared = .11]. Figure 20 displays this relationship graphically. Results indicated that subjects scoring better on this task had smaller (that is, more positive) N170 effect sizes.

IX. DISCUSSION

Grand-average ERPs - Early Components

Statistical analyses of the early components elicited by our manipulations in the visual wordform portion of the study did not indicate a replication of Kim & Lai's (2011) finding of an enhanced P100 component for contextually supported but misspelled words. Since the majority of stimuli that we employed in this portion were taken directly from their original design, it may be that our decision to change the manipulation of the critical word from a pseudoword ("The backpacker finally found a campsite and set up his *tont* before dark") to a phonologically illegal misspelling (" … his *tnet* before dark") may have contributed to all conditions showing equivalent P100 amplitudes. This may be because early stages of visual word recognition may depend on the phonological acceptability of a word in order for early interactions with higher-level semantic processes to occur. In other words,

visual words may need to meet some basic level of orthographic and phonological regularity (i.e. must contain onset syllables that are allowable in the language) in order to be initially considered as a semantically informative piece of input. If a string of characters do not meet these basic requirements, feed-forward interaction with "higher" levels of processing may be disrupted, which is what Kim & Lai (2011) suggested as a possible source of their effect. Differences in the ERP between control and experimental conditions did not manifest until the N170 time window (approximately 200 ms after word onset), and differences within experimental conditions did not manifest until at least 120 ms later. This suggests that subjects were not sensitive to the contextual supportedness of the misspellings until much later in the course of processing, and that it took them at least 320 ms to discern any possible semantic differences between critical misspellings in the three experimental conditions. Since it is arguable that phonologically illegal character strings flagrantly violate what a reader's expectations for how a wordform should look, it is plausible that the early processing involved with the perception of these stimuli varies in fundamental ways from how phonologically acceptable words or pseudowords are handled. In particular, if words are first "parsed" into syllable-like constituents, as some theories (e.g. Holcomb & Granger 2009) of reading hold, then processing of misspellings of the sort we employed here will differ from processing of phonologically legal strings very early on in the visual stream, perhaps too early for semantic factors to influence. In any case, we are currently conducting a replication of Kim & Lai's (2011) original manipulations in order to gain a better understanding of what factors are necessary to observe semantic influences on early-stage visual word-form perception.

Grand Average ERPs - Later Components

Although we failed to see significant between-condition differences in our early-stage visual ERPs, our manipulations of contextual supportedness did result in a clear parametric modulation of the

P600 component, with the degree of contextual supportedness being positively correlated with effect size. This appears to be yet another case in which a semantic factor can modulate the P600, which has been traditionally supported with structural processing. The effect is also consistent with an interpretation of the P600 as reflecting so-called "re-analysis" operations, which are posited to modulate the P600 in inverse proportion to the ease with which anomalous input can be "repaired" to be consistent with the context. To illustrate this idea, consider again the results of Kim & Sikos (2011), contrasted stimuli of the following types (reprinted below for convenience):

(KS1) The hearty meal was *devoured...* *Original Control Condition*

(KS2) The hearty meal was *devouring...* *Original Semantic Attraction Condition / Single Edit Repair*

(KS3) The hearty meal would *devour* *Multiple Edit Repair*

Sentences of the type (KS2) above were similar in design to the original semantic attraction condition of Kim & Osterhout (2005), and were intended to create a situation in which the inflection of the verb ("devouring") might simply be changed to -ed to create a well-formed and meaningful passive sentence (e.g. KS1). Sentences in the multiple-edit condition (KS3) could not be repaired as easily, as converting them to a well-formed passive would involve at least two operations: changing the both modal verb (from "would" to "was") and the inflection on the main verb (from "devour" to "devoured"). While both experimental conditions elicited P600 effects, the P600 elicited by the single -repair sentences was much larger in amplitude.

We argue that our P600 results can be interpreted in a similar light. Misspellings in the contextually supported condition are easily repaired towards the *most expected word* given the context (i.e. "tnet" to "tent"), and therefore elicit the largest P600 effect. Misspellings in the mid-support

condition are also repairable to words that are related, but not strictly predicted by, the foregoing context, and therefore elicit an effect size that is between the two other experimental conditions. Finally, misspellings in the no-support condition were designed *not* to resemble any words supported by the context, and therefore elicit the smallest P600 effect.

It is interesting to note that the size of the P600 effect is *greater* when the repair appears to be easier, which seems to run counter to the intuition that more complex repairs should recruit more resources and therefore elicit larger effects. Although it is beyond the scope of this paper to fully address this apparent contradiction, it may be that the re-analysis operations that underpin the P600 are only engaged to the extent that anomalous input resembles words that are consistent with both the preceding semantic context *and* structural cues. More generally, it may be that the tendency of the comprehension system to launch into a process of reanalysis is dependent on the existence of a *conflict* between competing representations of a phrase or sentence, i.e. when input cues conflict strongly with a comprehender's expectations for that input. Since the language system is not 'delusional', that is, will disregard a prediction in the face of input that obviously disconfirms it, we should expect that the ability of predictions to compete with actual input should be limited to cases in which the input could reasonably be re-analyzed as consistent with those predictions. Thus, in the single-edit repair condition of Kim & Sikos (2011), comprehenders are likely to have a strong expectation for a word similar in meaning to “devour”, but with a passive-compatible inflection (given both the auxiliary “was” and the inactivity of the subject noun “meal”). Since the semantic content of the actual input (“devouring”) partially confirms expectations, but the morphology does not, the system's prediction can compete with the input and force a re-analysis of the anomalous word, presumably to consider the possibility that it *does* in fact conform to expectation. Stimuli in the multiple-edit repair condition, however, do not engender the same expectations within the subject, since the modal verb “would” strongly predicts a following verb in infinitival form (regardless of the semantic context), a prediction which is ultimately

born out.

In terms of the present data, a similar story might be told: in all conditions, the context is highly predictive of a narrow range of possible words, and the ability of expectations to compete with the input should be a function of how close the input matches those expectations. If the input closely matches what is predicted (which holds for the supported and, to a lesser extent, mid-supported conditions), the expectations of the reader are more able to force a re-analysis of the input to make it compatible with the reader's representation of the context up until that point. However, if the input grossly deviates from expectations (which holds for the unsupported misspellings), then the prediction is falsified "without a fight".

It is worth noting that this story of "conflict" is a slight re-interpretation of Kim & Sikos (2011), in which they interpret P600 effect sizes as reflecting the ease with which re-analysis operations can succeed. Their interpretation puts syntax in conflict with semantics, in the sense that re-analysis operations reflected by their particular P600 effect in the single-repair condition overturn syntactic cues in favor of a more semantically plausible interpretation. It should be noted, however, that a conflict between syntax and semantics cannot directly account for the P600s that were elicited by our misspellings, since orthographic representations are not typically considered as syntactic cues. Instead, we are suggesting that it is conflict between expectations and actual input which triggers re-analysis, and that this could occur in any context in which anomalous input might be re-cast as consistent with the system's predictions or representation of the preceding discourse.

Individual Differences in Early ERPs

In the spirit of exploration, we were curious to see whether any of our behavioral measures related to the early ERP components that we measured in response to our misspelled target words.

Since there was little variance in the P100 component in response to our manipulations, we wondered if the average individual amplitude of the P100 generated in response to all items in this condition could be predicted from any of the behavioral measures we collected. Indeed, we found that subjects scoring higher on our vocabulary task tended to have greater P100 amplitudes. While our study was not designed to investigate this relationship, it is nevertheless worth a brief discussion.

Since many vocabulary tests (such as the one we administered) evaluate the participant's knowledge of infrequent English words which are usually only deployed in writing, the tasks can be thought of us an indirect measure of reading experience and/or ability. Thus, in broad terms, it should not be surprising that an ERP component associated with the recognition of visual words be related to performance on vocabulary tasks. That said, there are a number of mechanisms which might account for the relationship between visual wordform recognition and vocabulary size. One simple, but speculative, explanation is that frequent readers are more capable of sustained visual attention while reading for comprehension. Since the P100 component has been found to be a reliable index of selective visual attention (e.g. Luck, Heinze, Mangun & Hillyard 1990), it would seem reasonable to conclude that frequent and/or skilled readers may have been attending to our stimuli more intensely than their lower-performing peers.

X. CONCLUSION

In the foregoing experiments we examined, primarily, how individual differences in WM abilities can shape comprehension processes by revealing a relationship between how flexibly a subject can update verbal WM and how that in turn affects physiological measures known to index aspects of semantic and structural processing. We found that subjects scoring higher on our measure of verbal working memory updating tended to show P600 effects to outright animacy violations, while low

updaters tended to show only the predicted N400 effect. We concluded that limitations on WM updating constrained low performing subjects to consider alternative analyses only when the semantic content of the words in the sentence supported such analyses (i.e. in our semantic attraction condition). Better updaters, on the other hand, were able to consider alternative analyses even when the semantic context did not suggest any.

Although our manipulations in the VWF portion of the study failed to reveal any clear early effects of semantic process on early stages of word recognition, we were nevertheless able to observe clean, parametric differences in later ERP components that were elicited by misspelled words preceded by varying levels of contextual constraint. We interpreted these results as indicating that the amplitude of the P600 component is highly sensitive to the degree of match between a reader's contextually-generated expectations for a given visual word-form and the actual visual input. We suggested that the size of the P600 effects that we (as well as others) have observed is crucially dependent on the ability of the listener to “make sense” of the anomalous input, and perhaps more specifically the ability of predicted representations to influence bottom-up visual processing.

Although the exploratory nature of this work must be kept in mind when evaluating its results, we believe that Individual Differences (ID) aches to the study of language comprehension can contribute important perspective to ongoing psycholinguistic research. By examining how more general cognitive constraints can impinge on the comprehension process, we can learn how the many dimensions of linguistic input can be combined in unique ways across individuals to yield converging interpretations of an utterance. Although the individual variability we observed in our ERP measures suggests that the comprehension process can unfold in different ways depending on an individual's cognitive profile, further work must be completed to more clearly specify how more general cognitive constraints can shape language processes. Below we outline some suggestions for continued development of the ID approach to understanding language comprehension by detailing alternative

independent and dependent measures that might be employed towards further examination of some of the issues that we have raised here.

Furthermore, ID analyses of language-related ERPs can help determine if a given effect is shown consistently across individuals within a grand-average, or if it is subject to some amount of variability. For instance, in our study we observed that the so-called "semantic P600", as well as the N400 effect elicited by our no-attraction items, was fairly stable across individuals, indicating that it could reflect invariant processing operations that occur in these types of linguistic situations. When effects are more variable, however, it may indicate the situation in question is capable of being processed in many different ways, and thus represents an opportunity for examining what other cognitive factors influence the type of processing operation that is executed by one type of individual versus another.

In terms of future research that examines how more general cognitive faculties interact with language processes, we may consider measuring other cognitive functions besides WM when pursuing work that relates behavioral measures to physiological ones like ERPs. For instance, we may consider including tasks that stress other so-called “executive functions”, such as shifting and inhibition. It is possible that shifting, or the ability to flexibly switch consideration between different tasks or mental sets (c.f. Miyake, Friedman, Emerson, Witzki, Howerter & Wager 2000) may play a role in comprehension processes, particularly when listeners or readers are faced with situations in which they must weigh different types of linguistic cues when computing an interpretation. Inclusion of a shifting task in future research might further help to illuminate why some of our subjects appeared consider both semantic *and* structural cues with equal emphasis, while others (such as the low-updaters) seemed to rely more heavily on what the structural cues dictated.

Inclusion of inhibition measures in future studies may also be helpful. In another study of individual differences in language-related ERPs, Bornkessel et al (2004) suggested that high-span

readers sometimes employ their comparatively larger WM capacities towards the inhibition of competing, but dispreferred, sentential representations. Inclusion of an inhibition measure might allow us to test this hypothesis more directly by examining whether or not it is a better predictor of P600 and N400 effect sizes than the reading span task.

XI. REFERENCES

Bornkessel, I., Fiebach, C. Friederici A. (2004) On the cost of syntactic ambiguity in human language comprehension: an individual differences approach, Cognitive Brain Research, Volume 21, Issue 1, September 2004, pp 11-21

Caplan, D. Waters, G. S. (1999) Verbal working memory and sentence comprehension. Behavioral and Brain Sciences. VOL 22; Number 1, pp 77-94

Daneman, M., Carpenter P. (1980) Individual differences in working memory and reading, Journal of Verbal Learning and Verbal Behavior, Volume 19, Issue 4, August 1980, pp 450-466

Dewart, MH. (1979) Role of animate and inanimate nouns in determining sentence voice. British Journal of Psychology, Issue 70, pp. 135–141.

Judd, C. M., & McClelland, G. H. (1989). Data analysis: A model-comparison approach. San Diego, CA: Harcourt Brace Jovanovich.

Kim, A., Lai, V. (2011) Rapid Interactions between Lexical Semantic and Word Form Analysis during Word Recognition in Context: Evidence from ERPs. Journal of Cognitive Neuroscience, Vol 23 Issues 1-9

Kim, A. & Osterhout L. (2005) The independence of combinatory semantic processing: Evidence from event-related potentials Journal of Memory and Language, Vol. 52, No. 2. (February 2005), pp. 205-225.

Kim, A. & Sikos, L. (2011) Conflict and surrender during sentence processing: An ERP study of syntax-semantics interaction. Brain and Language, Vol 118, Issues 1-2. pp 15-22.

Kutas, M., Hillyard S.A. (1980) Reading senseless sentences: brain potentials reflect semantic incongruity. Science (New York, N.Y.), Vol. 207, No. 4427. (11 January 1980), pp. 203-205.

Luck, SJ., Heinze, HJ., Mangun, GR., & Hillyard, SA. (1990) Visual event-related potentials index focussed attention within bilateral stimulus arrays. II. Functional dissociation of P1 and N1 components. Electroencephalography and Clinical Neurophysiology, 75, 528–542.

Master, P. (1991) Active verbs with inanimate subjects in scientific prose. English for Specific Purposes, 10 (1991), pp. 15–33.

Miyake, A., Friedman, N., Emerson, M., Witzki, A., Howerter, A. (2000) The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. Cognitive Psychology, Vol 41(1), Aug 2000, 49-100.

Nakano H., Saron C., Swaab TY. (2010) Speech and span: working memory capacity impacts the use of

animacy but not of world knowledge during spoken sentence comprehension. Journal of Cognitive Neuroscience Vol 22 Issue 12 pp 2886-98

Osterhout, L., Holcomb, P. (1992) Event-related brain potentials elicited by syntactic anomaly. Journal of Memory and Language, Volume 31, Issue 6, pp 785-806

Semlitsch HV., Anderer, P., Schuster, P., Presslich, O. (1986) A solution for reliable and valid reduction of ocular artifacts, applied to the P300 ERP. Psychophysiology, Vol 23, Issue 6, pp 695-703

Trueswell, JC. & Tanenhaus, MK. (1994) Toward a lexicalist framework for constraintbased syntactic ambiguity resolution, C. Clifton, L. Frazier, K. Rayner, Editors , Perspectives on sentence processing, Erlbaum, Hillsdale, NJ, pp. 155–180.